统计学习能力测量的信度研究：基于模态、材料特征和测试任务的证据

于文勃[1] 亓鹤潼[1] 王天琳[2] 梁丹丹[1,3]

1.南京师范大学文学院

2.纽约州立大学奥尔巴尼分校教育学院

3.中国科学技术大学语言交叉研究中心

摘要：统计学习能力常作为自变量预测个体语言能力的发展，但以组间差异为目标编制的实验任务普遍信度不高，难以满足心理测量学的基本要求。本研究使用混合长度的目标结构合成学习材料，并且使用了二选一迫选和熟悉度评分两种测试任务，本研究设计了听觉语音和视觉图形两种模态，计算了测验的内部一致性系数和分半信度。结果发现使用混合长度目标结构编制实验任务的信度普遍高于以往研究；同时，视觉模态下的信度要高于听觉模态；前者迫选任务的信度也要高于熟悉度评分任务。综上，本研究推荐在视觉模态下，使用混合长度的目标结构合成学习材料，同时以迫选任务考察被试的统计学习能力。

关键词：统计学习能力，信度，迫选任务，熟悉度评分任务

1 引言

以基本认知能力作为自变量预测其他高级认知能力是心理学中一个常见的研究思路，随着对研究技术和研究方法要求的提高，心理学界越来越关注实验范式（任务）的科学性和测量的准确性问题，已有学者指出使用传统认知实验对某项认知能力进行测量时，信度往往不高，难以满足心理测量学对信效度[1]的要求（Hedge et al., 2018）。在语言心理学领域，统计学习被认为是和口语词切分、词汇语义习得等语言习得相关的基本认知能力（Estes et al., 2007， 2015; Newport, 2016; Saffran & Kirkham, 2018; Bogaerts et al., 2020; Siegelman, 2020; 徐贵平等，2020）。统计学习能力的传统测验范式在设计时是以组间差异视角为出发点的，关注的是被试因变量的组平均值是否高于某一个标准（如单样本 T 检验）或某几个被试组的组平均值是否有显著差异（如独立样本 T 检验），这一类范式在应用到以个体差异视角（如回归或相关）的研究时，就会出现测验信度不高、能力评估不稳定的问题，进而导致一些研究发现了统计学习能力能够预测语言发展的水平，一些研究则发现二者之间没有显著关系的现象（如 Lammertink et al., 2020）。在统计学习领域，已有几篇文章从信效度角度出发，质疑以传统统计学习任务结果为自变量预测语言发展结果的可靠性（Siegelman et al., 2017; Siegelman et al., 2018a）。本研究从测验信度出发，对传统测量方式进行修改，并验证其有

[1] 信度是效度的基础，如果信度不高，效度则不可能高，所以，信度高是效度高的必要不充分条件。

效性，一方面希望对统计学习能力的评估提供帮助，另一方面希望学界更加关注认知实验的信效度问题。

统计学习指个体能够从外界输入的时间信息和空间信息中发现统计规律并以此学习新事物的过程（Saffran et al., 1996; Frost et al., 2020; 于文勃等，2021（a），2021（b）; Isbilen & Christiansen, 2022），最经典的统计学习任务来自于 Saffran（1996）等的文章，采用的是学习-测试范式，学习材料由 4 个等长度的目标词（如图 1，每个目标词由三个音节组成，每个大写字母代表一个音节）按照伪随机的方式拼接而成，每个目标词在学习材料中出现 45 次。测试阶段，主试分别向被试播放目标词和跨界词，通过对比被试的注意时间来判断被试是否实现了统计学习。后续针对幼儿和成人的实验沿用了学习阶段的材料，在测试阶段多使用迫选任务（详见 Isbilen & Christiansen, 2022），其中每个试次包括一个目标词和一个跨界词（如 CJK）或非词[2]（BHE），要求被试选择出组成学习材料的基本单位。由于跨界词是两个目标词之间的转换之处，因此被认为是词边界，记忆效果不强；而目标词内部的音节始终相连，音节组合关系更加紧密，记忆效果也就牢固。统计实验结果时，如果被试组别的迫选正确率显著高于 0.5，那么就认为出现了学习效应。

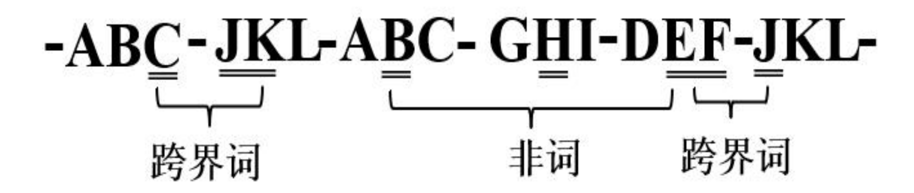


图 1 语音统计学习材料示意图

近年来，学界在个体差异研究视角下，开始将被试在迫选任务中的正确个数作为统计学习能力的指标，进而预测典型发展儿童语言发展和解释多种障碍儿童的认知和语言表现的原因（Erickson et al., 2016; Kidd & Arciuli, 2016; von Koss Torkildsen, 2019; Kidd et al., 2020; Isbilen et al., 2022）。虽然这一个体差异视角下的研究得出了不少显著的结论，但依托组间差异视角的实验任务信度较低，本文总结了部分统计学习任务的信度结果（表 1），可以看到大多数都难以满足心理测量学对能力测量信度的最低标准：0.8（Nunnally & Bernstein, 1994）。Siegelman（2017）提出组间差异视角研究范式面临的两个问题：（1）测试任务中试次太少（通常为 16 个）；（2）测试阶段始终使用跨界词和目标词进行配对比较，难度一致。这两个因素共同导致被试得分的变异较小，依托于相关分析而得到的测验信度也就较低。

[2] 统计学习中非词没有固定的组成方式，只要是在学习材料中前后不相连的音节组合就可以称作非词。

此外，迫选任务中为了平衡顺序效应，同一个选项（包括目标结构和跨界结构）还要多次出现，不仅降低了敏感性还会对信度产生影响。

一些统计学习研究在实验任务的介绍中也会报告内部一致性系数，但很少见到系统比较不同模态、不同任务下信度指标差异的研究，学界更没有一个针对统计学习能力相对完善的测验方案。Arnon（2020）使用经典的实验范式，分别计算了成人和儿童完成多种统计学习任务的信度指标，发现成人被试的信度指标达到中等程度，但儿童被试的信度很低，和心理测量学的要求相距甚远。Siegelman（2017）则对以往视觉统计学习任务进行修改，虽然信度指标得到大幅提升，但测验时长大大增加，还包括了多种试题形式，不利于在婴幼儿和发展障碍儿童身上广泛使用。

表 1 部分统计学习实验的信度

作者	模态	试次数	样本量	α系数	分半信度	重测信度
Siegelman,2015	视觉	32	76			0.58
Siegelman,2017	视觉	42	62	0.88	0.72--0.9	
Siegelman,2018	语音	42	55	0.42		
Siegelman,2018	视觉	36	200	0.84		
Siegelman,2018	视觉	36	200	0.78		
Siegelman,2018	语音	36	200	0.54		
Siegelman,2018	语音	36	200	0.59		
Tong,2019	视觉	32	35	0.56		
Arnon,2020[3]	语音	25	52	0.57	0.18--0.63	0.61
Arnon,2020	视觉	25	52	0.83	0.55--0.83	0.45
Kidd,2020	语音	32	37	-0.04		
Kidd,2020	语音	32	37	-0.05		
Witteloostuijn,2021	视觉	24	50		0.55—0.8	
Witteloostuijn,2021	视觉	16	50		0.67--0.85	

除了以上两个问题以外，还有研究指出统计学习过于理想化的前提也是一个影响信度的因素。统计学习以“白板假说”为前提（如 Elazar et al., 2022），假设被试在学习任务前未接触过人工语言，测试阶段所表现出的学习效应均来自于学习阶段。但事实上，语音统计学习中学习的音节（组合）在被试的母语中存留痕迹[4]，每一个被试的语言经验不同，对迫选试次的判断也就存在异质性，测验的内部一致性系数自然会较低。目前来看，以音节为材料的统计学习研究最多，但也存在其他材料，如音调（Saffran et al., 1999）、声音（Siegelman

[3] 本文以成人为被试，只总结了 Arnon（2020）原文中成人被试的结果。

[4] 语音统计学习要求学习材料为无意义语音材料，在印欧语系下只能保证音节组合没有对应的词，在汉藏语系下只能保证音节没有对应的字（词），但印欧语系中音节组合可能是有意义词语的部分组合，汉藏语系中，也可能在方言中出现或音段结构容易产生联想。

et al., 2018a）和图形（Siegelman et al., 2018b），详见元分析文章 Frost（2020）等。有研究认为视觉图形不容易受到被试经验的影响，也更容易满足“白板假说”，信度更高，更应该作为统计学习能力的测量任务（Siegelman et al., 2018a）。

基于以上分析可以看出，统计学习能力的测量陷入困境，制约了探讨统计学习和语言能力关系的研究。本研究主要从试次的难度差异、测验任务和材料模态等方面对传统实验任务进行改进，具体来看，不同于以往研究使用等长度的目标结构合成学习材料，本研究以不同长度的目标结构合成学习材料，这是因为不同长度的目标结构会对应不同的转换概率和记忆表征，可以丰富试次的难度差异，提高被试得分的变异；同时，还可以避免被试产生节奏期待夸大实验效应（Hoch et al., 2013）。其次，近年来一些学者使用熟悉度评分任务作为统计学习能力的测试任务（Batterink et al., 2015），即要求被试对目标结构、跨界结构和非结构的熟悉程度进行评分，这一任务可以避免同样的选项在迫选试次中反复出现而降低试次的敏感性，本研究也将检验熟悉度评分任务的信度指标，为测量任务提供备选。第三，我们还分别设计了视觉和听觉模态的任务，以比较不同模态下统计学习测验的信度。

综上，本研究立足于找到一个更为有效的统计学习能力测评方案。首先，通过修改材料特征和使用熟悉度评分测试两种方式获得被试得分的更大变异性；同时依据“白板假说”设置了视觉图形模态的统计学习任务。我们预期测验的信度会得到明显提升，尤其是在使用熟悉度评分任务的视觉图形模态下。考虑到被试量、实验任务复杂程度等因素，本研究并未在材料特征这一变量中设置等长度目标结构的水平，而是全部以混合长度目标结构合成人工语言。在统计检验过程中，本研究未使用如 T 检验等参数检验，而是使用类似元分析的方式，主要对比八个测验和以往研究测验（表 1）的信度差异，并以此判断测验方案的优劣。

2 方法

2.1 被试

共有 159 名被试参与实验，男性被试 47 名，被试年龄范围 18 ~ 27 岁，所有被试母语均为汉语普通话。参加听觉语音学习材料 A 的被试 41 人，学习材料 B 的被试 39 人，参加视觉图形学习材料 A 的被试 40 人，学习材料 B 的 39 人[5]。实验前，被试签署知情同意书，实验结束后被试获取少量报酬，本研究经过校伦理委员会审查（× × 2022060023 和 × × 202302010）。

2.2 实验设计

[5] 由于部分被试操作失误，听觉语音模态下迫选任务和熟悉度评分任务的有效被试量均为 80 人，视觉图形模态下迫选任务的有效被试量为 74，熟悉度评分任务的有效被试量为 77。

本研究仍旧采用学习-测试范式，实验设计是 2（模态：视觉图形，听觉语音）×2（测试任务：熟悉度评分任务，迫选任务）×2（对照材料：学习材料 A，学习材料 B）的三因素的混合实验设计；其中，模态和对照材料是被试间变量，对照材料中的学习材料 A 中的目标结构是学习材料 B 中的跨界结构，反之，学习材料 B 中的目标结构是学习材料 A 中的跨界结构，这一设置可以保证实验效应不是来自于特殊的材料组合。测试任务是被试内变量。被试随机分配到四个被试间水平中，一半被试先完成熟悉度评分任务，一半被试先完成迫选任务，从而平衡测试任务的顺序效应。本研究的因变量为测验任务的信度，包括克伦巴赫α系数和分半信度的区间范围。

2.3 材料和实验程序

2.3.1 听觉语音材料

学习材料的编制参考（于文勃等，2021b）的研究。在汉语普通话音节库中选择符合发音规则的 10 个音节，包括 CV（C 代表辅音，V 代表元音）和 CVV 两种形式，这两种音节形式是普通话中最常见的结构。为了避免声调承载的统计信息影响结果，所有音节均为第一声，而且这 10 个音节没有对应的汉字，尽量避免被试进行联想。由一名女性普通话母语者在专业录音室进行录音，采样率为 44100Hz。通过将目标音节放置在两个音节之间来排除录音人对目标音节的重度或明显的停顿，例如录音人一次性产出音节串 nve1-ruo1-gei1，其中只有 ruo1 为目标音节。随后通过 Praat 软件分离目标音节，并对其持续时间（300ms）、平均音高（266Hz）和强度（70dB）进行归一化（http://www.praat.org/）。使用 10 个音节随机组合成两组无意义的目标词，分别为两个两音节目标词和两个三音节目标词，通过 Praat 脚本合成学习材料 A 和学习材料 B，要求同一个目标词不能连续出现两次，而且其后出现其他词语的可能性相等（1/3）。学习材料 A 和 B 中每个目标词重复出现 120 次，总计 480 个词，呈现时长为 6 分钟。

在强迫选择任务中，每个试次由 1 个目标词和 1 个跨界词组成，目标词和跨界词的长度相等以避免词长对被试的选择产生影响。每个目标词和两个跨界词进行迫选，一半的试次中目标词先出现，另一半中跨界词先出现，以此来平衡顺序效应。另外，三音节迫选对和两音节迫选对各 8 个，共计 16 个试次。在熟悉度评分任务中，被试对目标词、跨界词和非词的熟悉程度进行七点评分。两个测试任务试次的呈现顺序均为随机。三类词具体如表 2。

表 2 语音音节任务的三类词材料

学习材料版本	目标词	跨界词	非词
语音学习材料 A	nueruote diafolai	tediafo lainueruo	nuemeilai diasete

	remei	meirou	refo
	rouse	sere	rouruo
	tediafo	diafolai	terouruo
语音学习材料 B	lainueruo	nueruote	lairefo
	meirou	rouse	meinue
	sere	remei	sedia

2.3.2 视觉图形材料

图形材料的选择参考（Siegelman 等，2018b）的研究，选择 10 个无意义图形组成学习材料。为了保证获得稳定的学习效应，学习阶段中每个图形呈现 800ms，随后出现 100ms 的空白，接着出现下一个图形，即 SOA 为 900ms。人工材料编制的原则同听觉语音模态，每个目标图形组合呈现 28 次，迫选任务和熟悉度评分任务的设计也和音节任务保持一致，无论是目标结构、跨界结构还是非结构都是一个整体呈现在电脑屏幕上要求被试进行迫选或评分，三类图形组合如表 3。

表 3 视觉图形任务的三类图形材料

图形材料版本	目标结构	跨界结构	非结构
视觉学习材料 A			
视觉学习材料 B			

2.3.3 实验程序

实验程序由 E-prime 呈现，听觉语音条件下被试佩戴耳机完成，电脑音量固定为 30%。两种模态下实验程序都包括练习实验和正式实验，练习实验前由主试讲解实验要求和指导语，学习阶段播放 5s 中的学习材料，随后完成迫选任务和熟悉度评分任务。练习实验中的材料在正式实验中不会出现。语音模态下实验任务大约需要 15 分钟完成，视觉图形任务大约需要 10 分钟。实验流程图如图 2。实验材料、数据和代码见：https://github.com/****** /reliability-of-SL。

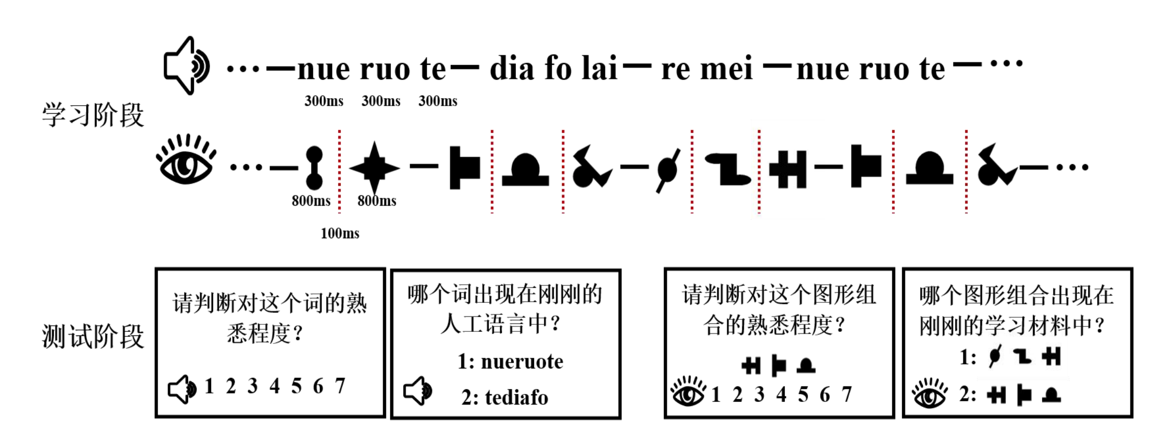


图 2 实验流程图

3 结果

采用 R 语言（4.3.1）对数据分析，克伦巴赫α系数和分半信度均使用 psych 包中的 reliability 函数进行计算，数据结果见表 3。

表 3 克伦巴赫α系数和分半信度结果

模态	听觉语音模态				视觉图形模态			
测试任务	熟悉度评分任务		迫选任务		熟悉度评分任务		迫选任务	
对照	A	B	A	B	A	B	A	B
α 系数	0.659	0.754	0.651	0.679	0.719	0.736	0.863	0.740
分半信度	0.406	0.594	0.199	0.332	0.561	0.397	0.590	0.379
	0.823	0.906	0.898	0.885	0.860	0.881	0.973	0.943

结合表 1 和表 3 的数据，本研究测验任务的信度指标和其他研究结果的关系如图 3，可以看出以混合长度目标结构合成的统计学习任务，信度指标和以往研究相当或更好。此外，我们还统计了被试在所有任务上的学习效应。结果发现，在听觉模态下的迫选任务中，被试

的正确率显著高于随机水平（t（79）= 5.180，$p < 0.001$，95%CI：[0.070，0.160]，$d = 0.580$）；同时，熟悉度评分任务中，被试对目标词的评分显著高于跨界词（t（79）= 4.570，$p < 0.001$，95%CI：[0.330，0.840]，$d = 0.510$），对目标词的熟悉度也显著高于非词（t（79）= 10.640，$p < 0.001$，95%CI：[1.250，1.830]，$d = 1.540$）。在视觉图形模态下的迫选任务中，被试的正确率显著高于随机水平（t（73）= 5.930，$p < 0.001$，95%CI：[0.100，0.200]，$d = 0.680$）；同时，熟悉度评分任务中，被试对目标结构的评分显著高于跨界结构（t（76）= 6.580，$p < 0.001$，95%CI：[0.840，1.57]，$d = 1.200$），对目标结构的评分也显著高于非结构（t（76）= 12.770，$p < 0.001$，95%CI：[2.290，3.140]，$d = 2.720$）。这些结果一致说明被试在本研究的多个实验任务中都表现出了稳定的学习效应。最后，我们还发现被试完成熟悉度评分任务和迫选任务结果的相关性，听觉语音模态下，两种任务被试得分的相关系数 $r = 0.460, p < 0.001$，95%CI：[0.254，0.609]，视觉图形模态下，两种任务被试得分的相关系数 $r = 0.450, p < 0.001$，95%CI:[0.233, 0.611]。

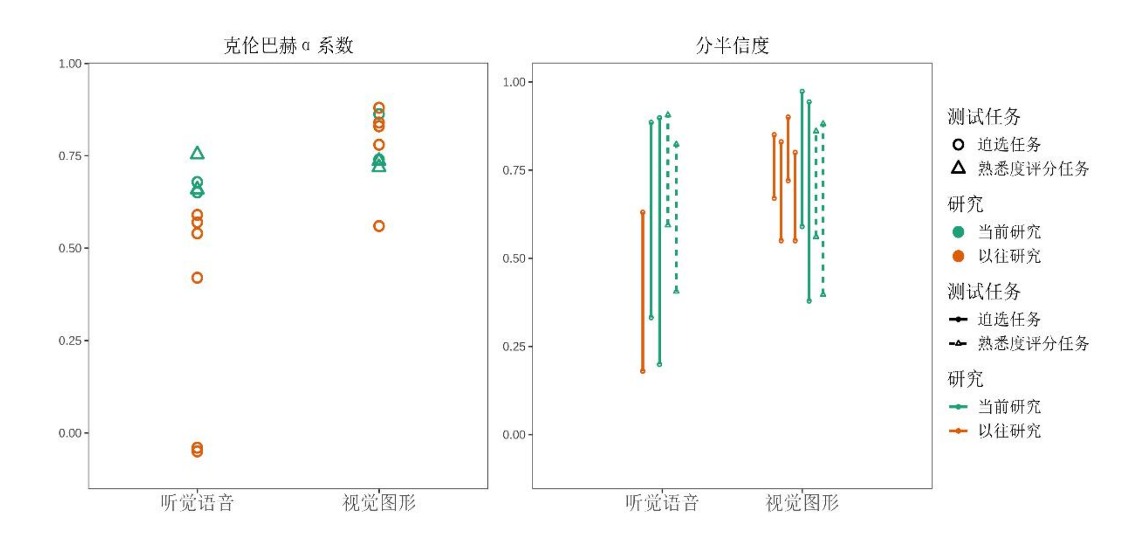


图 3 统计学习能力测验信度汇总

4 讨论

在探讨统计学习能力和个体语言发展的关系时，尤其要关注统计学习能力的评估方式，以往研究的信度指标不够理想，引起了很多争论和探讨。本研究在学习材料上使用了混合长度的目标结构，并且对比了迫选任务和熟悉度评分任务的信度结果；此外，还纳入了视听模态的对比。结果发现，视觉模态下统计学习测验的α系数较高，基本达到心理测量学的要求，同时分半信度区间也更为理想，尤其是迫选任务的信度指标好于熟悉度评分任务。

4.1 混合长度学习材料对信度的影响

以听觉语音模态为例，本研究在听觉语音模态中使用三音节和两音节目标词合成学习材料，在视觉图形模态中使用三连图形和两连图形合成学习材料。依据统计学习的记忆组块机制（Perruchet, 2019; Isbilen et al., 2020），被试在测试任务中根据学习阶段的记忆表征进行选择，三音节目标词和三音节跨界词有共同的音节组合，同时仅在一个跨界处有明显的词边界特征，因此记忆表征强度差异较小，被试选择难度较大；相反，两音节目标词和两音节跨界词没有共同的音节组合，但在一个跨界处有词边界特征，记忆表征差异较大，被试选择时难度也较小。所以，以混合长度目标结构合成学习材料时，迫选试次难度区分更细致。在熟悉度评分任务中由于添加了非词结构，试次包括三（两）音节目标结构、三（两）音节跨界结构和三（两）音节非结构，非结构在学习阶段没有出现过，是最容易判断的试次，因此，熟悉度评分任务中试次难度差异更大，也更容易获得变异较大的得分。实验的信度指标也符合我们的预期，无论使用熟悉度评分任务还是迫选任务，混合长度统计学习任务的测验信度都与以往研究持平或高于以往研究。

4.2 学习材料模态对信度的影响

以克伦巴赫α系数为信度指标时，视觉图形模态下的统计学习任务信度均高于 0.70，尤其以迫选任务作为测验任务时，信度达到 0.863 和 0.740，（基本）符合心理测量学的标准（Nunnally & Bernstein, 1994）；相反，听觉语音模态下的测验信度不高，有三个条件都低于 0.7。在计算分半信度时，本研究只报告了区间，不过即便这样，视觉模态下信度的上限和下限都高于语音模态，而且变异范围更窄，所以使用视觉材料有助于提高统计学习任务的信度。

本研究的结果和 Siegelman（2018a）等人的观点一致：相比于语音材料，视觉统计学习较少受到被试个体语言经验的影响，被试间的判断一致性更高。Siegelman（2017，2018a）等修订的视觉任务中，学习阶段每个目标结构只呈现 24 次，测试阶段被试需要完成 42 个试次，包括熟悉度评分和图形补全两类任务，同时干扰选项也更多。他们研究中视觉任务的克伦巴赫 α 系数分别为 0.84 和 0.78，和本研究基本一致。但在我们的研究中，仍旧使用了较为常见的二选一强迫选择任务，试次数也只有 16 个。考虑到统计学习任务很多时候会应用在儿童被试身上，所以实验任务是否简短、有效也是测验的一个重要指标。所以相对来说，本研究的任务兼顾了测量学要求和实验效益，也更可能应用到幼儿和儿童研究中。

虽然个别研究指出听觉统计学习能力能够预测个体的阅读技能或解释阅读障碍儿童的阅读障碍表现（Gabay, Thiessen, & Holt, 2015; Qi et al., 2019），但大多数研究使用视觉材料作为统计学习材料（如 Tong et al., 2019，详见 Lee et al., 2022）。尤其是从学理上来说，视

觉图形和汉字这种象形文字会有更大的关联。本研究则发现以视觉图形作为学习材料对统计学习能力的评估更为稳定, 这一结果将进一步推动统计学习能力和汉语儿童阅读技能关系的研究。尤其是在普通话背景下，符合发音规则但无意义的音节很少，以第一声为例仅有 20 个左右，这也给语音统计学习任务的材料编制带来很大困扰。因此，结合心理测量学要求和实验材料的可操作性，我们更推荐使用视觉模态任务作为统计学习能力的考察方式。最后，本研究还计算了不同模态内部迫选任务和熟悉度评分测验成绩的相关性。从两种测验的结果来看，模态内部测验成绩存在显著相关，语音模态内的相关性已经得到很多研究的支持（Erickson et al., 2016），但视觉模态下不同任务的相关性研究还鲜有见到，未来还需要更多的研究进行深入探讨。

4.3 测验任务类型对测验信度的影响

本研究除了使用迫选任务外，还使用了熟悉度评分任务作为统计学习的测试任务，这是因为迫选任务中为了平衡选项出现的顺序，每个选项都要重复出现多次，所以迫选任务的结果包含了被试在学习阶段的学习效应和迫选过程中的二次学习效应, 这一平衡策略会降低测验的敏感性。不过，本研究结果发现在视觉图形模态下，熟悉度评分任务的α系数普遍低于迫选任务，分半信度的区间也要更大、更低，因此从心理测量学角度来看，迫选任务是对统计学习能力更好的评估方式。

4.4 不足和展望

本研究有一些不足之处。关于测验形式的问题上，一些研究从构想效度的角度对不同测验任务所考察的内容进行分析，认为迫选任务和熟悉度评分任务都属于反思类任务，不仅考察了个体捕捉统计信息的能力, 还包括了元认知的能力 (Ordin & Polyanskaya, 2021; Isbilen & Christiansen, 2022），所测量的统计学习能力并不纯粹，后续研究应该从更多方面衡量不同的实验任务。关于学习材料参数细节上，被试的学习效应不仅受到学习材料概率信息的影响，还和材料的呈现时长、呈现次数有关（Bogaerts et al., 2016）。未来研究如果以婴幼儿为目标被试，还应该考虑这些因素的影响。

5 结论

为满足心理测量学的基本要求，本研究对统计学习能力的测量方案进行修改，发现在视觉图形模态下，使用混合长度的目标结构合成学习材料，以迫选任务作为测验任务的组合方式，能够获得较为稳定的信度指标。

参考文献

徐贵平, 范若琳, 金花. (2020) .统计学习的认知神经机制及其与语言关系.*心理科学进展, 28*(09): 1525-1538.

于文勃，王璐，程幸悦，王天琳，张晶晶，梁丹丹.（2021a）.语言经验对概率词切分的影响.*心理科学进展,* 29(05)：787-795.

于文勃，王璐，瞿邢芳，王天琳，张晶晶，梁丹丹.（2021b）. 转换概率和词长期待对语音统计学习的影响. *心理学报，53*(06)：565-574.

Arnon, I. (2020). Do current statistical learning tasks capture stable individual differences in children? an investigation of task reliability across modality. *Behavior Research Methods52* :68-81.

Batterink, L. J., Reber, P. J., Neville, H. J., & Paller, K. A. (2015). Implicit and explicit contributions to statistical learning. *Journal of Memory and Language, 83*, 62-78.

此文献为补充的参考文献：

Bogaerts, L., Siegelman, N., & Frost, R. (2016). Splitting the variance of statistical learning performance: A parametric investigation of exposure duration and transitional probabilities. *Psychonomic bulletin & review*, *23*, 1250-1256.

Bogaerts, L., Frost, R., & Christiansen, M. H. (2020). Integrating statistical learning into cognitive science. *Journal of Memory and Language,115*,104167.

Elazar, A., Alhama, R. G., Bogaerts, L., Siegelman, N., Baus, C., & Frost, R. (2022). When the “Tabula” is anything but “Rasa:” What determines performance in the auditory statistical learning task?. *Cognitive Science, 4*6(2), e13102.

Erickson, L. C., Kaschak, M. P., ED Thiessen, & Berry, C. (2016). Individual differences in statistical learning: conceptual and measurement issues. *Collabra, 2*(1), 14.

Estes, K. G., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can Infants Map Meaning to Newly Segmented Words? Statistical Segmentation and Word Learning. *Psychological Science, 18*(3), 254–260.

Estes, K. G., Gluck, C. W., & Bastos, C. (2015). Flexibility in statistical word segmentation: finding words in foreign speech. *Language Learning and Development, 11*(3), 252-269.

Frost, R., Armstrong. B. C., & Christiansen. M. H. (2020). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin, 145*(12), 1128-1153.

Gabay, Y., Thiessen, E. D., & Holt, L. L. (2015). Impaired statistical learning in developmental dyslexia. *Journal of Speech, Language, and Hearing Research, 58*(3), 934-945.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior research methods, 50*, 1166-1186.

Hoch, L., Tyler, M. D., & Tillmann, B. (2013). Regularity of unit length boosts statistical learning in verbal and

nonverbal artificial languages. *Psychonomic Bulletin & Review, 20*(1), 142–147.

Isbilen, E. S., McCauley, S. M., Kidd, E., & Christiansen, M. H. (2020). Statistically induced chunking recall: a memory-based approach to statistical learning. *Cognitive Science, 44*(7).

Isbilen, E. S., & Christiansen, M. H. (2022). Statistical Learning of Language: A Meta‐Analysis Into 25 Years of Research. *Cognitive Science, 46*(9), e13198.

Isbilen, E. S., McCauley, S. M., & Christiansen, M. H. (2022). Individual differences in artificial and natural language statistical learning. *Cognition, 225* (2022) 105123.

Kidd, E., & Arciuli, J. (2016). Individual differences in statistical learning predict children's comprehension of syntax. *Child development*, 87(1), 184-193.

Kidd, E., Arciuli, J., Christiansen, M. H., Isbilen, E. S., Revius, K., & Smithson, M. (2020). Measuring children's auditory statistical learning via serial recall. *Journal of Experimental Child Psychology, 200*, 104964.

Lammertink, I., Boersma, P., Rispens, J., & Wijnen, F. (2020). Visual statistical learning in children with and without DLD and its relation to literacy in children with DLD. *Reading and Writing, 33*(6), 1557-1589.

Lee, S. M. K., Cui, Y., & Tong, S. X. (2022). Toward a model of statistical learning and reading: Evidence from a meta-analysis. *Review of Educational Research, 92*(4), 651-691.

Newport, E. L. (2016). Statistical language learning: Computational, maturational, and linguistic constraints. *Language and Cognition,* 8(3), 447−461.

Nunnally, J., & Bernstein, I. (1994). Psychometric theory (3rd ed.). New York: McGraw-Hill.

Ordin, M., & Polyanskaya, L. (2021). The role of metacognition in recognition of the content of statistical learning. *Psychonomic Bulletin & Review, 28*, 333-340.

Perruchet, P. (2019). What mechanisms underlie implicit statistical learning? Transitional probabilities versus chunks in language learning. *Topics in cognitive science, 11*(3), 520-535.

Qi, Z., Sanchez Araujo, Y., Georgan, W. C., Gabrieli, J. D., & Arciuli, J. (2019). Hearing matters more than seeing: A cross-modality study of statistical learning and reading ability. *Scientific Studies of Reading, 23*(1), 101-115.

Saffran, J. R., & Kirkham, N. Z. (2018). Infant Statistical Learning. *Annual Review of Psychology*, 69(1), 181–203.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274, 1926−1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70*(1), 27-52.

Siegelman, N. (2020). Statistical learning abilities and their relation to language. *Language and Linguistics Compass, (14)*, e12365.

Siegelman, N., Bogaerts, L., & Frost, R. (2017). Measuring individual differences in statistical learning: current pitfalls and possible solutions. *Behavior Research Methods, 49*(2), 1-15.

Siegelman, N., Bogaerts, L., Elazar, A., Arciuli, J., & Frost, R. (2018a). Linguistic entrenchment: Prior knowledge impacts statistical learning performance. *Cognition, 177*, 198-213.

Siegelman, N., Bogaerts, L., Kronenfeld, O., & Frost, R. (2018b). Redefining "Learning" in Statistical Learning: What Does an Online Measure Reveal About the Assimilation of Visual Regularities? *Cognitive Science, 42* (S3), 692-727.

Tong, X., Leung, W. W. S., & Tong, X. (2019). Visual statistical learning and orthographic awareness in Chinese children with and without developmental dyslexia. *Research in developmental disabilities, 92*, 103443.

von Koss Torkildsen, J., Arciuli, J & Ona Bø Wie. (2019). Individual differences in statistical learning predict children's reading ability in a semi-transparent orthography. *Learning and Individual Differences, 69*(2019), 60-68.

Reliability Study of Statistical Learning Ability Measurement: Evidence from Modality, Material and Task

Wenbo Yu[1], Hetong Qi[1], Tianlin Wang[2], Dandan Liang[1,3]

1.School of Chinese Language and Culture, Nanjing Normal University

2. School of Education, University at Albany, State University of New York

3.Interdisciplinary Research Center for Linguistic Science, University of Science and Technology of China

Abstract: Research has considered statistical learning (SL) as a fundamental learning mechanism in cognition, for which individuals rely on the statistical regularities from visual and verbal input during information processing. Take the verbal SL task as an example, participants are first exposed to a nonsensical artificial language or visual sequence for 5~10 mins and then asked to finish a 2 alternative forced choice task (2AFC). Accuracy on each trial is coded in a dichotomous manner, with 0 for incorrect and 1 for correct, and aggregated across participants to generate the mean accuracy of the group. If it is higher than chance level, it is assumed that learning has occurred. This research perspective is called the perspective of inter group differences.

However, this index is the result from the perspective of inter-group differences, which is suitable for judging whether the test group exhibits statistical learning effects, but not measuring

the relationship between SL ability and other cognitive ability. Some researchers criticized the low reliability of SL tasks and suggested that the task results are not psychometrically satisfactory. In the current study, we aimed to put forward a modified SL task that is relatively more comprehensive. Two aspects of traditional tasks have been modified; one is that we constructed learning materials with mixed-lengths targets, and another is that we employed a familiarity rating task to measure learning outcomes in addition to the 2AFC task.

A total of 159 participants took part in our experiment. Two types of reliability Cronbach's alpha coefficient and split-half reliability were computed with the *reliability* function in R. The results of this study are divided into three aspects. Firstly, the index of two types of reliability in the current study are better than previous studies. This indicates that the learning materials we constructed with mixed-length nonsensical words exhibit some advantages in reliability. Secondly, the results revealed that both the Cronbach's alpha coefficient and split-half reliability of statistical learning tasks in the visual modality were higher than those in the auditory modality, which is consistent with the opinion of Siegelman (2018a). Then, the reliability of forced-choice tasks in the visual modality was higher than that of familiarity rating tasks, suggesting the results obtained from 2AFC task are more stable and consistent across participants.

The current study explored the task in measuring SL ability, underscoring the importance of using mixed-length learning materials and suggest employing visual stimuli in assessing statistical learning abilities in addition to the traditional utilization of forced-choice tasks during the testing phase. Future studies should not only focus on designing brief SL tasks for children and language disorder population that align with psychometric standards, but also rethink the cognitive mechanism underlying various SL task.
Key words: statistical learning; 2-alternative forced choice task; familiarity rating task; reliability